Power Evaluation of Disease Clustering Tests

International Journal of Health Geographics

Research

Power evaluation of disease clustering tests

Changhong Song*[1] and Martin Kulldorff[1,2]

Address: [1]Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, U.S.A and [2]Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care,133 Brookline Avenue, 6th Floor, Boston, MA 02215, USA

Email: Changhong Song* - changhon@stat.uconn.edu; Martin Kulldorff - martin_kulldorff@hms.harvard.edu

* Corresponding author

Published: 19 December 2003

International Journal of Health Geographics 2003, **2**:9

Received: 30 October 2003

Accepted: 19 December 2003

This article is available from: http://www.ij-healthgeographics.com/content/2/1/9

Abstract

Background: Many different test statistics have been proposed to test for spatial clustering. Some of these statistics have been widely used in various applications. In this paper, we use an existing collection of 1,220,000 simulated benchmark data, generated under 51 different clustering models, to compare the statistical power of several disease clustering tests. These tests are Besag-Newell's *R*, Cuzick-Edwards' *k*-Nearest Neighbors (*k*-NN), the spatial scan statistic, Tango's Maximized Excess Events Test (MEET), Swartz' entropy test, Whittemore's test, Moran's *I* and a modification of Moran's *I*.

Results: Except for Moran's *I* and Whittemore's test, all other tests have good power for detecting some kind of clustering. The spatial scan statistic is good at detecting localized clusters. Tango's MEET is good at detecting global clustering. With appropriate choice of parameter, Besag-Newell's *R* and Cuzick-Edwards' *k*-NN also perform well.

Conclusion: The power varies greatly for different test statistics and alternative clustering models. Consideration of the power is important before we decide which test statistic to use.

Background

A large number of tests for spatial randomness that adjust for an uneven background population have been proposed. Such test statistics are used to test whether or not the geographical distribution of disease is random. They are also used in many other areas such as genetics, geomorphology and ecology [1-6].

When we use these test statistics, it is important to know whether they have good power. There have been some studies comparing such test statistics [7-14], but there have been few simultaneous comparisons of three or more tests. When evaluating tests for spatial randomness, the best way is to compare them using the same simulated data sets.

For our study, we use existing benchmark data [10], simulated from the female population in the Northeastern United States, to evaluate the power of different test statistics for various kinds of clusters.

Previous studies have shown that the spatial scan statistic has good power in detecting hot spot clusters, and Tango's MEET has good power in detecting global clustering [10]. We compare the power of these two test statistics with six additional tests: Besag-Newell' *R*, Cuzick-Edwards' *k*-NN, Swartz' entropy test, Whittemore's test, Moran's *I* and a

modified version of Moran's *I*. These tests are selected for different reasons. Some tests are widely used, such as Moran's I and Cuzick-Edwards' *k*-NN. Most of them are published in well reputed statistics journals.

Methods

Benchmark data sets

The benchmark data sets are based on the 1990 female population in the 245 counties and county equivalents in the Northeastern United States, consisting of the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland and the District of Columbia. Each county is represented by a centroid coordinate. The data is available at 'http://www.commed.uchc.edu/biostat/datasets/'. The benchmark data and how it was generated has been described in detail elsewhere [10]. We provide a brief summary here.

Under the null hypothesis of no clustering, 100,000 random data sets were generated by randomly allocating 600 cases to the various counties, with probabilities proportional to the county population. The null data is used to estimate the critical values, which is the cut-off point for the significance. Two kinds of clustering models were evaluated, hot spot clusters and global chain clustering.

Hot spot clusters

Hot spot clusters were generated by setting the relative risk in some counties to be larger than 1. Three different sets of local clusters are constructed in a rural, urban and mixed area respectively. Within each of these three sets, there are five different sized clusters with 1, 2, 4, 8 and 16 counties respectively. The center of the rural cluster is Grand Isle County in Vermont. The center of the mixed cluster is Allegheny County (Pittsburgh) in Pennsylvania. The center of the urban cluster is New York county (Manhattan) in New York. The relative risks and counties included in each cluster are listed in Table 1.

In order to evaluate how the disease clustering tests perform when there are multiple hot-spot clusters, the benchmark data also include 15 alternate models with two clusters and 5 models with three clusters by using different combinations of the original clusters. In a model, all clusters had the same number of counties.

Global chain clustering

In the global chain clustering model, every county has the same expected number of cases under the null and alternative hypothesis. The counties are tied together sequentially on a chain that passes through the centroid of each county exactly once, after which it reconnects with the first county on the chain, forming a Hamiltonian cycle. A map of the Hamiltonian cycle used has previously been published [10].

Table 1: The hot spot clusters

	Counties	Counties included	$E[c\|H_0]$	$E[c\|H_A]$	Relative risk
Rural clusters	1	Grand Isle, VT	0.05	10	192.89
	2	above + Franklin, VT	0.46	12	27.03
	4	above + Clinton, NY, Chittenden, VT	2.69	18	7.05
	8	above + Lamoille, VT, Washington, VT, Essex, NY, Addison, VT	4.16	22	5.35
	16	above + Orleans, VT, Franklin, NY, Caledonia, VT, Orange, VT, Essex, VT, Rutland, VT, Warren, NY, Windsor, VT	7.32	28	3.90
Mixed clusters	1	Allegheny, PA	14.43	39	2.85
	2	above + Washington, PA	16.41	42	2.70
	4	above + Beaver, PA, Westmoreland, PA	22.52	51	2.40
	8	above + Butler, PA, Armstrong, PA, Lawrence, PA, Fayette, PA	27.47	58	2.24
	16	above + Greene, PA, Indiana, PA, Clarion, PA, Mercer, PA, Somerset, PA, Venango, PA, Cambria, PA, Jefferson, PA	34.22	67	2.10
Urban clusters	1	New York, NY	15.97	42	2.73
	2	above + Hudson, NJ	21.78	50	2.43
	4	above + Bronx, NY, Kings, NY	59.99	100	1.81
	8	above + Queens, NY, Bergen, NJ, Essex, NJ, Richmond, NY	101.96	150	1.63
	16	above + Union, NJ. Nassau, NY, Passaic, NJ, Rockland, NY, Westchester, NY, Morris, NJ, Middlesex, NJ, Monmouth, NJ	154.94	209	1.53

To generate clusters, a certain number of cases are first located randomly on the map, according to the null hypothesis. These original cases then generate other new cases close by. If each original case generates one additional case, it is called twins. If two additional cases are generated, it is called triplets.

A total of 26 chain clustering models are constructed with the distance between the twins (triplets) along the chain being either constant or exponentially distributed with different means. If the distance is zero, the twins (triplets) are in the same county. The chain does not imply that the disease itself spreads around the chain, just that twins and triplets cases are located in either of the two directions, as defined by the chain.

Test statistics

Notation

Denote c_i as the number of cases in county i, n_i as the population size of county i, C as the total number of cases, N as the total population size, H as the total number of counties and d_{ij} as the distance between county i and j.

Let $D_{j(i)}$ be the total number of cases in county i and its j closest neighbors, and let $U_{j(i)}$ be the population size in county i and its j closest neighbors.

Besag-Newell's R

Besag-Newell's R statistic [15] has been used to study leukemia in upstate New York [16]. The test statistic is defined as $R = \sum_{i=1}^{H} c_i I(P(M_i \le m_i) < 0.05)$, where M_i is a random variable denoting the minimum number of counties needed to have at least k cases in county i and its M_i closest neighboring counties, m_i is the observed value of M_i, that is, $m_i = min\{j : (D_{j(i)} + 1) \ge k\}$. k is a parameter set by the user. Usually, a large k is more sensitive to large clusters and a small k is more sensitive to small clusters. I is the indicator function with value 1 when $P(M_i \le m_i) <$ 0.05 and 0 otherwise. $P(M_i \le \mathrm{m}_i)$ is calculated by

$$P(M_i \le m_i) = 1 - P(M_i > m_i) = 1 - \sum_{s=0}^{k-1} e^{-U_{m_i(i)}\frac{C}{N}} (U_{m_i(i)}\frac{C}{N})^s / s!.$$

The null hypothesis of no clustering is rejected when R is large.

Cuzick-Edwards' k-NN

Cuzick-Edwards' k-NN (k-Nearest Neighbors) test [17] has been widely used, for example for leukemias and lymphomas among young people in New Zealand [18] and the association of Ixodes pacificus and quine granulocytic ehrlichiosis in California [19].

This test statistic was originally designed for point data, but can easily be adapted for aggregated data. The test statistic is defined as

$$T_k = \sum_i c_i g_i^k,$$

where k is a parameter chosen by the user and for each county i, g_i^k denotes the number of k nearest neighbors which are cases. To be more precise, $g_i^k = D_{(h\text{-}1)(i)} + t_{h(i)}$ where h is decided so that $U_{(h\text{-}1)(i)} \le k$, $U_{h(i)} > k$ and $t_{h(i)} = \frac{k - U_{(h-1)(i)}}{U_{h(i)} - U_{(h-1)(i)}} c_h$. $U_{(\text{-}1)(i)}$ is defined as 0.

The null hypothesis of no clustering is rejected when T_k is large.

The spatial scan statistic

The spatial scan statistic [20] has among other things been used to study human granulocytic ehrlichiosis near Lyme in Connecticut [21], soft-tissue sarcoma and non-Hodgkin's lymphoma clusters with high dioxin emission levels [22], childhood mortality in rural Burkina Faso [23], bovine tuberculoisis in Argentina [24] and Toxoplasma gondii infection of southeast sea, otters [25].

The spatial scan statistic imposes a circular window on the map and lets the circle centroid move across the study region. For any given position of the centroid, the radius of the window is changed continuously to take any value between zero and some upper limit.

Let $L_{j(i)}$ be the likelihood under the alternate hypothesis that there is a cluster in county i and its j closest neighbors, and let L_0 be the likelihood under the null hypothesis. It can then be shown that

$$\frac{L_{j(i)}}{L_0} = (\frac{D_{j(i)}}{U_{j(i)}\frac{C}{N}})^{D_{j(i)}} (\frac{C - D_{j(i)}}{C - U_{j(i)}\frac{C}{N}})^{C - D_{j(i)}}.$$

As this likelihood ratio is maximized over all circles, it identifies the one that constitutes the most likely cluster. The test statistic is

$$T = \max_{i,j} \frac{L_{j(i)}}{L_0} I(D_{j(i)} > \frac{U_{j(i)}}{N} C),$$

where I is the indicator function with value 1 when $D_{j(i)} > \frac{U_{j(i)}}{N} C$ and 0 otherwise. The null hypothesis of no clustering is rejected when T is large.

Tango's Maximized Excess Events Test (MEET)

For a given parameter λ, the Excess Events Test statistic [11] is defined as

$$EET(\lambda) = \sum_i \sum_j e^{-4d_{ij}^2/\lambda^2}(c_i - n_i\frac{C}{N})(c_j - n_j\frac{C}{N}).$$

The choice of λ relates to the geographical scale of clustering. Large λ makes the test sensitive to geographically large clusters, while small λ will make the test more sensitive to small clusters.

To be able to detect clustering irrespectively of its geographical scale, Tango [12] proposed the Maximized Excess Events Test (MEET)

$$MEET = \min_{0\le\lambda\le U} P(EET(\lambda) > eet(\lambda) \mid H_0, \lambda),$$

where $eet(\lambda)$ is the observed value of the Excess Events Test statistic conditioning on λ, and U is the upper limit on λ. Practical implementation of the test uses 'line search' by discretization on λ, and the MEET statistic is evaluated using Monte Carlo hypothesis testing [26].

The null hypothesis of no clustering is rejected when the test statistic is small.

Swartz' entropy test

Swartz [27] proposed a test for spatial randomness based on the concept of entropy. The test statistic is defined [28] as

$$T = ln(C!) + ln((N-C)!) - \sum_i (ln(c_i!) + ln((n_i - c_i)!)).$$

The null hypothesis of no clustering is rejected when T is small.

Moran's I

Moran's I [29] was originally proposed to analyze continuous data. Subsequently, this statistic has also been used to analyze count data, such as Lyme disease in New York State [30] and cancer incidence in Canada [31].

The Moran' I statistic is defined as

$$I = \frac{\sum_{i=1}^{H}\sum_{j=i+1}^{H}(r_i - \bar{r})(r_j - \bar{r})a_{ij}}{\sum_{i=1}^{H}(r_i - \bar{r})^2}.$$

where $\bar{r} = \frac{1}{H}\sum_{i=1}^{H} r_i$, $r_i = \frac{c_i}{n_i}$ and

$$a_{ij} = \begin{cases} 1 & \text{if county i and j are neighbors.} \\ 0 & \text{if county i and j are not neighbors.} \end{cases}$$

We also consider a modified version of Moran's I:

$$I_{mod} = \sum_{i=1}^{H}\sum_{j=i+1}^{H}(r_i - \bar{r})(r_j - \bar{r})a_{ij},$$

In both cases, we reject the null hypothesis of no clustering when I is large.

Whittemore's test

Whittemore et al. [32] proposed the statistic

$$T = \frac{1}{2}\sum_i \sum_j d_{ij}c_i c_j.$$

We reject the null hypothesis of no clustering when T is small.

Power calculation

For Besag-Newell's R, Cuzick-Edwards' k-NN. Swartz' entropy test, Moran's I and Whittemore's test, the power estimate is calculated using C++ code written by the author. For the spatial scan statistic and Tango's MEET, the power estimates are obtained from the paper by Kulldorff et al.[10].

Results

Hot spot clusters

Table 2 shows the estimated power of the test statistics in detecting the hot spot clusters. For each type of hot spot cluster, the highest power is highlighted. The spatial scan statistic has good power in detecting all three kinds of hot spot clusters: rural, mixed and urban clusters, and it performs best for detecting rural clusters. Tango's MEET performs best for the urban clusters, but not very well for the rural clusters.

With the right choice of parameter, Besag-Newell's R has the best power for detecting mixed clusters, but its strength is very sensitive to choice of parameter. The power of Cuzick-Edwards' k-NN also depends on the parameter. It has good power in detecting all three kinds of hot spot clusters with the right choice of parameter. The choice of parameter depends on the size of the cluster. Usually, for large clusters, large parameter values perform better, while for small clusters, small parameter values are better.

Swartz' entropy test has good power in detecting the rural clusters, but not very good for mixed or urban clusters. Moran's I can detect the rural clusters except for the cluster with only one county. The modified Moran's I has similar performance to Moran's I, but it performs better for the rural clusters, especially for the cluster with one county. Whittemore's test does not perform as well as the other test statistics in detecting hot spot clusters.

Table 2: Power of the test statistics for the hot spot clusters.

		Besag-Newell's *R*			Cuzick-Edwards' *k*-NN			Spatial Scan Statistic	Tango's MEET	Swartz' Entropy Test	Moran's		Whittemore's Test
		k = 6	12	30	100K	500K	1000K				*I*	I_{mod}	
Rural (edge)	1	0.707	0.388	0.089	0.752	0.168	0.038	**0.998**	0.196	0.939	0.000	0.315	0.010
	2	0.792	0.466	0.074	0.810	0.199	0.049	**0.991**	0.221	0.804	0.743	0.793	0.006
	4	0.839	0.754	0.239	0.874	0.425	0.109	**0.973**	0.229	0.607	0.449	0.505	0.004
	8	0.830	0.854	0.309	0.851	0.540	0.157	**0.971**	0.213	0.639	0.752	0.814	0.002
	16	0.821	0.880	0.505	0.758	0.621	0.247	**0.969**	0.229	0.706	0.715	0.806	0.004
Mixed (corner)	1	0.037	0.023	**0.983**	0.648	0.919	0.899	0.936	0.925	0.270	0.053	0.045	0.000
	2	0.129	0.024	**0.989**	0.655	0.886	0.913	0.939	0.896	0.289	0.059	0.051	0.000
	4	0.157	0.095	**0.980**	0.645	0.822	0.931	0.937	0.838	0.269	0.078	0.061	0.000
	8	0.217	0.222	**0.956**	0.608	0.777	0.903	0.941	0.817	0.291	0.130	0.099	0.000
	16	0.293	0.284	0.914	0.598	0.715	0.838	**0.949**	0.832	0.354	0.193	0.165	0.000
Urban (central)	1	0.037	0.027	**0.952**	0.627	0.856	0.893	0.922	0.941	0.264	0.049	0.045	0.296
	2	0.033	0.214	0.819	0.587	0.786	**0.937**	0.903	0.920	0.245	0.056	0.049	0.334
	4	0.026	0.049	0.190	0.378	0.684	0.864	0.892	**0.961**	0.119	0.052	0.043	0.579
	8	0.022	0.022	0.459	0.292	0.637	0.817	0.913	**0.983**	0.078	0.061	0.043	0.758
	16	0.015	0.059	0.368	0.257	0.648	0.795	0.926	**0.986**	0.047	0.069	0.045	0.887
Rural and Mixed	1	0.624	0.270	0.981	0.956	0.943	0.860	**1.000**	0.964	0.975	0.000	0.310	0.000
	2	0.803	0.356	0.987	0.969	0.926	0.891	**0.999**	0.952	0.923	0.727	0.780	0.000
	4	0.841	0.739	0.983	0.977	0.930	0.929	**0.997**	0.930	0.813	0.460	0.508	0.000
	8	0.867	0.906	0.972	0.973	0.939	0.916	**0.996**	0.931	0.849	0.767	0.823	0.000
	16	0.857	0.938	0.970	0.949	0.939	0.891	**0.996**	0.941	0.914	0.729	0.810	0.000
Mixed and Urban	1	0.020	0.012	**0.999**	0.929	0.997	0.998	0.987	0.998	0.545	0.049	0.041	0.009
	2	0.084	0.132	0.995	0.918	0.990	**0.998**	0.984	0.995	0.499	0.057	0.045	0.012
	4	0.082	0.069	0.946	0.807	0.962	**0.996**	0.966	0.991	0.303	0.080	0.048	0.034
	8	0.107	0.114	0.915	0.710	0.935	0.987	0.954	**0.990**	0.222	0.136	0.073	0.070
	16	0.120	0.167	0.803	0.616	0.897	0.969	0.935	**0.984**	0.199	0.212	0.135	0.138
Rural and Urban	1	0.619	0.272	0.954	0.949	0.902	0.868	**1.000**	0.970	0.974	0.000	0.309	0.096
	2	0.709	0.665	0.823	0.955	0.854	0.919	**0.999**	0.962	0.909	0.712	0.771	0.097
	4	0.731	0.644	0.261	0.947	0.863	0.855	**0.992**	0.971	0.671	0.436	0.472	0.206
	8	0.676	0.689	0.546	0.911	0.879	0.826	**0.991**	0.977	0.602	0.726	0.770	0.365
	16	0.591	0.725	0.521	0.803	0.892	0.834	**0.987**	0.975	0.561	0.659	0.726	0.562
Rural, Mixed and Urban	1	0.541	0.185	0.998	0.992	0.998	0.994	**1.000**	0.999	0.991	0.000	0.291	0.002
	2	0.735	0.565	0.993	0.994	0.994	0.997	**1.000**	0.998	0.960	0.697	0.755	0.001
	4	0.735	0.611	0.949	0.987	0.984	0.995	**0.996**	0.994	0.799	0.433	0.456	0.003
	8	0.728	0.759	0.922	0.972	0.980	0.984	**0.992**	0.989	0.759	0.730	0.769	0.007
	16	0.642	0.766	0.840	0.909	0.962	0.962	0.977	**0.983**	0.744	0.672	0.737	0.023

All the test statistics have good power for multiple hot spot clusters except Whittemore's test and Moran's I. The spatial scan statistic, Tango's MEET and Cuzick-Edwards' *k*-NN perform very well in detecting multiple clusters.

Global chain clustering

Table 3 shows the estimated power of the test statistics for global chain clustering. The highest power for each type of global clustering is highlighted. Note that as the distance between the cases increases, there is less clustering in the data, and all tests have lower power.

For most alternative models, Tango's MEET has the highest power. The spatial scan statistic performs well, but not as well as Tango's MEET. Swartz' entropy test is good when the distance is small, but the power decreases very quickly as the distance increases. Besag-Newell's R, Moran's I,

Table 3: Power of the test statistics for the global chain clustering.

		Twins											
		Besag-Newell's *R*			Cuzick-Edwards' *k*-NN			Spatial Scan Statistic	Tango's MEET	Swartz' Entropy Test	Moran's		Whittemore's Test
		k = 6	12	30	100K	500K	1000K				*I*	I_{mod}	
No distance	0.00	0.477	0.491	0.423	**1.000**	0.925	0.728	0.791	0.990	0.999	0.049	0.136	0.132
Fixed Distance	0.005	0.076	0.242	0.332	0.488	**0.644**	0.570	0.392	0.624	0.357	0.116	0.101	0.128
	0.01	0.057	0.077	0.231	0.159	0.319	0.383	0.285	**0.406**	0.143	0.078	0.068	0.122
	0.02	0.060	0.060	0.118	0.077	0.107	0.154	0.194	**0.264**	0.079	0.056	0.054	0.116
	0.04	0.061	0.054	0.055	0.060	0.065	0.067	0.124	**0.174**	0.062	0.051	0.050	0.097
	0.08	0.056	0.054	0.050	0.059	0.059	0.058	0.080	**0.109**	0.059	0.050	0.051	0.073
	0.16	0.060	0.051	0.042	0.059	0.056	0.045	0.055	0.059	**0.060**	0.053	0.054	0.053
Exponential Distance	0.005	0.212	0.314	0.351	**0.820**	0.709	0.587	0.452	0.738	0.642	0.182	0.179	0.127
	0.01	0.140	0.210	0.274	0.534	0.525	0.466	0.351	**0.556**	0.386	0.133	0.121	0.122
	0.02	0.096	0.134	0.191	0.284	0.314	0.309	0.262	**0.378**	0.210	0.094	0.086	0.112
	0.04	0.076	0.097	0.121	0.144	0.171	0.184	0.185	**0.250**	0.127	0.071	0.067	0.102
	0.08	0.063	0.074	0.091	0.086	0.104	0.111	0.124	**0.166**	0.081	0.063	0.058	0.085
	0.16	0.059	0.063	0.061	0.062	0.070	0.074	0.080	**0.107**	0.064	0.054	0.053	0.071
		Triplets											
No distance	0.00	0.742	0.780	0.716	**1.000**	0.999	0.964	0.995	**1.000**	**1.000**	0.052	0.196	0.188
Fixed Distance	0.005	0.088	0.333	0.587	0.715	**0.885**	0.856	0.674	0.884	0.559	0.178	0.148	0.179
	0.01	0.064	0.092	0.368	0.228	0.470	0.587	0.491	**0.646**	0.202	0.102	0.087	0.171
	0.02	0.067	0.062	0.149	0.092	0.132	0.212	0.318	**0.430**	0.098	0.065	0.060	0.149
	0.04	0.060	0.063	0.057	0.076	0.079	0.084	0.189	**0.265**	0.072	0.054	0.053	0.118
	0.08	0.058	0.056	0.044	0.069	0.063	0.057	0.102	**0.141**	0.066	0.052	0.048	0.078
	0.16	0.060	0.057	0.035	**0.066**	0.053	0.044	0.046	0.050	0.064	0.053	0.053	0.043
Exponential Distance	0.005	0.315	0.524	0.629	**0.977**	0.939	0.867	0.762	0.960	0.884	0.317	0.314	0.176
	0.01	0.185	0.323	0.473	**0.786**	0.773	0.721	0.610	0.826	0.598	0.200	0.185	0.170
	0.02	0.118	0.184	0.303	0.438	0.489	0.490	0.436	**0.599**	0.315	0.127	0.117	0.154
	0.04	0.084	0.110	0.180	0.202	0.251	0.272	0.289	**0.390**	0.161	0.085	0.077	0.135
	0.08	0.073	0.075	0.099	0.110	0.118	0.139	0.171	**0.226**	0.098	0.062	0.060	0.102
	0.16	0.060	0.066	0.065	0.070	0.071	0.078	0.091	**0.115**	0.067	0.053	0.054	0.071

Whittemore' test are not very good at detecting global clustering.

With the right choice of parameter, Cuzick-Edwards' *k*-NN performs very well, especially for clustering with small distances. Large parameter values tend to detect clustering with large distance, while small parameter values perform better for clustering with small distance.

The performance of the test statistics for twins clustering with fixed and exponential distance is similar. All test statistics have better power in detecting triplet clustering since there is more clustering there.

Discussion

Of the evaluated test statistics, Besag-Newell's *R*, Cuzick-Edwards' *k*-NN, the spatial scan statistic, MEET, Whittemore's test are based on Euclidean distances. Moran's I is based on the adjacencies of counties. Swartz' entropy test does not use the spatial relationship among the counties.

The *M* statistic [33] proposed by Bonetti and Pagano is a nonparametric test that uses the iriterpoint distance distribution to study the spatial pattern of the data. The *M* statistic has also been evaluated using the same benchmark data [10]. The *M* statistic does well for mixed and urban clusters and has good power in detecting multiple clusters. Generally, it does not perform quite as well as the spatial scan statistic and MEET, but it is very competitive compared to the other tests.

Besag-Newell's *R* and Cuzick-Edwards' *k*-NN are good test statistics, but the power depends a lot on the parameter. Usually large parameter value can make the test statistic more sensitive to large clustering, whereas small parameter value can detect the small clustering better. So if we know the scale of clustering and choose a corresponding parameter, these two test statistics may have good power. In practice, we usually don't know the size of clustering. If we try different parameter values, that will cause multiple testing problems.

Sometimes we need to adjust the analysis for age or other covariate. All the test statistics considered here can incorporate such adjustment except Swartz' entropy test, although it can be modified to do so.

In terms of data resolution, Besag-Newell's R, Whittemore's test, Tango's MEET and Swartz' entropy test were originally proposed to analyze aggregated data, while Cuzick-Edwards' k-NN was proposed to analyze point data. The spatial scan statistic was proposed to analyze either aggregated or point data. Moran's I was designed for continuous data, but has been used extensively for aggregated count data as well. It is possible and maybe even likely that these test statistics may perform differently when applied to point data.

A strength of this power evaluation study is that the data is typical of epidemiological applications, and uses actual population and geographical data. The strength of the test statistics will depend not only on the alternative model though, but also on the spatial distribution of the areas and the population size in this area. A limit of the study is that the background population of the benchmark data is from only one particular region, the female population of Northeast United States. Under other alternate models and background population, some test statistics may perform better or worse.

Conclusion

The power varies greatly for different disease clustering test statistics. Consideration of the power is important before deciding which test statistic to use. If the size or scale of clustering is known, it is worth considering the use of Besag-Newell's R or Cuzick-Edwards' *k*-NN. If not, we feel confident recommending the spatial scan statistic for the detection of local clusters and use Tango's MEET for the general evaluation of clustering throughout the map. Other tests may be equally good or better for alternative models not considered in this paper.

List of abbreviations

k-NN: *k*-Nearest Neighbors.

MEET: Maximized Excess Events Test.

Authors' contributions

CH and MK jointly designed the study and chose the methods for evaluation. CH programmed the C++ code, carried out the power simulations and wrote the first draft of the manuscript. Both authors interpreted the results and wrote the final version of the paper.

Acknowledgements

This research was funded by NCI grant number RO1CA095979-01.

References

1. RuizGarcia M: **Genetic relationships among some new cat populations sampled in Europe: A spatial autocorrelation analysis.** *Journal of Genetics* 1997, **76:**1-24.
2. Gustine DL, Elwinger GF: **Spatiotemporal genetic structure within white clover populations in grazed swards.** *Crop Science* 2003, **43:**337-344.
3. Aubry P, Piegay H: **Spatial autocorrelation analysis in geomorphology: Definitions and tests.** *Geographic Phisique et Quaternaire* 2001, **55:**111-129.
4. Meirmans PG, Vlot EC, Den Nijs JCM, Menken SBJ: **Spatial ecological and genetic structure of a mixed population of sexual diploid and apomictic triploid dandelions.** *Journal of Evolutionary Biology* 2003, **16:**343-352.
5. Liebhold AM, Gurevitch J: **Integrating the statistical analysis of spatial data in ecology.** *Ecography* 2002, **25:**553-557.
6. Clark SA, Richardson BJ: **Spatial analysis of genetic variation as a rapid assessment tool in the conservation management of narrow-range endemics.** *Invertebrate Systematics* 2002, **16:**583-587.
7. Rogerson PA: **The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic.** *Geographical Analysis* 1999, **31:**130-147.
8. Kulldorff M, Nagarwalla N: **Spatial disease clusters: Detection and inference.** *Statistics in Medicine* 1995, **14:**799-810.
9. Oden N: **Adjusting Moran' I for population density.** *Statistics in Medicine* 1995, **14:**17-26.
10. Kulldorff M, Tango T, Park P: **Power comparisons for disease clustering tests.** *Computational Statistics and Data Analysis* 2003, **42:**665-684.
11. Tango T: **A class of tests for detecting 'general' and 'focused' clustering of rare diseases.** *Statistics in Medicine* 1995, **14:**2323-2334.
12. Tango T: **A test for spatial disease clustering adjusted for multiple testing.** *Statistics in Medicine* 2000, **19:**191-204.
13. Vach W: **Locally optimal tests on spatial clustering.** *in New Approaches in Classification and Data Analysis* Edited by: *Diday. Berlin, Springer-Verlag*; 1994:161-168.
14. Tango T: **Comparison of general tests for spatial clustering.** In *Disease Mapping and Risk Assessment for Public Health* Edited by: *Lawson, et al. London, Wiley*; 1999:111-117.
15. Besag J, Newell J: **The detection of clusters in rare diseases.** *Journal of the Royal Statistical Society* 1991, **A154:**143-155.
16. Waller LA, Turnbull BW, Clark LC, Nasca P: **Spatial pattern analyses to detect rare disease clusters.** In: *Case Studies in Biometry* Edited by: *Lange N, Ryan L, Billard L, Brillinger D, Conquest L, Greenhourse J. New York: John Wiley & Sons*; 1994:13-16.
17. Cuzick J, Edwards R: **Spatial clustering for inhomogeneous populations.** *Journal of the Royal Statistical Society* 1990, **B52:**73-104.
18. Dockerty JD, Sharples KJ, Borman B: **An assessment of spatial clustering of leukaemias and lymphomas among young people in New Zealand.** *Journal of Epidemiology and Community Health* 1999, **53:**154-8.
19. Vredevoe LK, Righter PJ, Madigan JE, Kimsey RB: **Association of Ixodes pacificus (Acari: Ixodidae) with the spatial and temporal distribution of equine granulocytic ehrlichiosis in California.** *Journal of Medical Entomology* 1999, **36:**551-561.
20. Kulldorff M: **A spatial scan statistic.** *Communications in Statistics: Theory and Methods* 1997, **26:**1481-1496.
21. Chaput EK, Meek JI, Heimer R: **Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut.** *Emerging Infectious Diseases* 2002, **8:**943-948.

22. Viel JF, Arveux P, Baverel J, Cahn JY: **Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels.** *American Journal of Epidemiology* 2000, **152:**13-19.
23. Sankoh OA, Ye Y, Sauerborn R, Muller O, Becher H: **Clustering of childhood mortality in rural Burkina Faso.** *International Journal of Epidemiology* 2001, **30:**485-492.
24. Perez AM, Ward MP, Torres P, Ritacco V: **Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in Argentina.** *Preventive Veterinary Medicine* 2002, **56:**63-74.
25. Miller MA, Gardner IA, Kreuder C, Paradies DM, Worcester KR, Jessup DA, Dodd E, Harris MD, Ames JA, Packham AE, Conrad PA: **Coastal freshwater runoff is a risk factor for Toxoplasma gondii infection of southern sea otters (Enhydra lutris nereis).** *International Journal for Parasitology* 2002, **32:**997-1006.
26. Dwass M: **Modified randomization tests for nonparametric hypotheses.** *Annals of Mathematical Statistics* 1957, **28:**181-187.
27. Swartz JB: **An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbors.** *Health and Place* 1998, **4:**67-77.
28. Kulldorff M: **Letter to the editor.** *Health and Place* 1999, **5:**313.
29. Moran PAP: **Notes on continuous stochastic phenomena.** *Biometrika* 1950, **37:**17-23.
30. Glavanakov S, White DJ, Caraco T, Lapenis A, Robinson GR, Szymanski BK, Maniatty WA: **Lyme disease in New York State: Spatial pattern at a regional scale.** *American Journal of Tropical Medicine and Hygiene* 2001, **65:**538-545.
31. Le ND, Marret LD, Roberson DL, Semenciw RM, Turner D, Walter SD: **Canadian Cancer Incidence Atlas.** *Canadian Government Publishing.* 1995.
32. Whittemore AS, Friend N, Brown BW, Holly EA: **A test to detect clusters of disease.** *Biometrika* 1987, **74:**631-635.
33. Bonetti M, Pagano M: **On detecting clustering.** *Proceedings of the Biometrics Section American Statistical Association* 2001:24-33.